Emergency Medicine Journal

Why sample size estimates?

Journal:	*Emergency Medicine Journal*
Manuscript ID	emermed-2018-207763.R1
Article Type:	Concepts
Date Submitted by the Author:	30-Jun-2018
Complete List of Authors:	Weber, Ellen; University of California San Francisco, Emergency Medicine Hoo, Zhe Hui; University of Sheffield, School of Health and Related Research (ScHARR); Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield Adult CF Centre
Keywords:	statistics, research, methods, research, clinical, publication, epidemiology

WHY SAMPLE SIZE ESTIMATES?
Understanding Sample Size Calculations and Statistical Significance

EMJ asks that authors include in their Methods section a sample size calculation. Authors may wonder why we require this, and readers may wonder why they just shouldn't skip over it. In this paper, we explain what information can be gleaned from the sample size estimate, and along the way, the true meaning of statistical significance.

Why should I calculate a sample size?

In every experiment, you are using a sample to make inferences about a much larger target population. The problem is that you don't know if your sample is truly representative of the entire population. Your sample may be younger or older than the rest, or have different illnesses. Clearly, the larger your sample, the more of this variation you will capture, the more representative your sample is likely to be and the more precise your estimates. But you don't have infinite resources. So the goal of a sample size calculation is to help to ensure we have enough subjects to take account of that underlying variation in determining whether, for example, treatment a is better than treatment b and if so, by how much?

In a descriptive study, you estimate the sample size based on the predicted proportion of patients that have the illness or the outcome. In a comparative study, you are looking to see if there is a difference in some outcome between two or more groups. The groups may differ by one or more aspects (e.g. one group is < 65 years and another is ≥ 65 years; or one group had training and another didn't, or one group got the drug and another didn't, as in a randomised trial.). The goal of a sample size calculation here is get a fairly precise estimate of the effect in each group, with reasonable assurance that you didn't get these results by chance. This is where we get into statistical significance.

Statistical Significance – What does it really mean?

Many people misinterpret what "statistically significant" means. Don't feel bad if you're one of them – the concept comes from the language of hypothesis testing, which is not something that is very intuitive to clinicians. Using the language of hypothesis testing, statistical significance means that the probability of rejecting the null hypothesis (that there is no effect of the intervention) when it is true is low. Translation: **Statistical significance means that result you got is unlikely to be due to chance. Importantly,** it does *not* mean that the numbers for each group are greatly different from each other, or that the difference is meaningful clinically.

Remember, you began with a sample of the target population. To be more certain of your result, you'd want to repeat the study on different samples in that same population. Statistical significance means that if you repeated the study multiple times, with different samples from the target population, you would get very similar results most of the time. Your answer is unlikely to be a fluke!

. Statistical significance is typically inferred from p values. To interpret a p-value, it is helpful to remember that the p value comes from the world of hypothesis testing, where the experiment is set out such that the goal is to determine if you can reject the null hypothesis. In a randomized study comparing two treatments for example, the null hypothesis is that there is no difference between the treatments. If you find a difference, then the p value tells you whether you have sufficient statistical evidence to reject the null hypothesis. A smaller p value implies a greater statistical incompatibility with the null hypothesis[1] In other words, the smaller the p value, the stronger the evidence for rejecting the null hypothesis.

Let's say you did an experiment to compare two variables and you obtained a set of values from that experiment. You then compare those variables using a statistical test, which generates a p-value. Let say that p-value is 0.03. The means the probability of obtaining that value (or one more extreme i.e. more different) by chance in that single experiment is 0.03. If you have specified a "significant" p-value as 0.05, it means the p-value you obtained is "statistically significant", i.e. unlikely you have obtained that set of results just by chance. But if you have specified a "significant" p-value as 0.01, that means the p-value you obtained is "not statistically significant".

In most studies, statistical significance (alpha) is set at .05. This means that the probably of getting the result you have gotten, or one more extreme, by chance is 5%. If you would like to be even more certain the result isn't due to chance, you set alpha at .01. Note that *this doesn't mean the magnitude of the effect is bigger*, only that you are more certain. In reality, when you do an experiment you calculate a p value, which may be greater or less than 0.05 or 0.01. If you have set your alpha at .05, and your p value is, say .03, you can call your results statistically significant; that is, the probability of getting the result you have gotten, or one more extreme, is within the range of chance you're willing to take.

Getting back to Sample Size –
Sample size for a comparative study is determined by the size of the effect you are looking for, and variation in the population and two other parameters – the level of statistical significance (alpha) and the power (1-B) . We've already said that by convention,0 .05 is usually chosen as the cut-off for statistical significance. Power is the likelihood that we won't reject the null hypothesis when it is actually false. Translation: Power is the likelihood of finding an effect when there is one. The higher the power (and lower the Beta), the greater the likelihood. When estimating the effect size, we may use parameters from other studies to guide us or we may choose an effect size we wish to see (below which, say, the cost of a treatment isn't worth it). The variation generally comes from other studies. These are of course estimates, and you don't know if the variation in your sample will be the same, and that can mean even if you reach the estimated sample size, and see an effect, it might not reach statistical significance.

It is always preferable to base your sample size on the desired effect size. Sometimes this may result in an unrealistically large sample size. In this case, researchers may adjust the significance level, or the power. It is tempting, but

probably a bad idea, to adjust the effect size. It might not seem obvious at first, but looking for larger treatment effects generally means needing smaller sample sizes: if the effect of a treatment is very large, you are likely to see it very early in your experiment, and this would suggest that you don't need a very large sample. It may be quite tempting to design a study this way to shorten recruitment time and lower your costs, but few things (in real life) have large effect sizes, so if you decide on too large an effect size, even if a clinically important differences exists, a "statistically significant" difference could not be detected with that sample size. Moreover, you've lost some researcher integrity here in no longer sticking with your hypothesized effect size.

Failing to show a significant difference when there might be one is called (in stats speak) a Type II error. Ultimately it boils down to a "too small "sample size, but this could be the result of a smaller effect or more variation than anticipated. This can happen in any experiment – a sample size estimate is just that, an estimate. But it's the best tool we have.

This leads to the corollary, which is that with a large enough sample size, even very small differences that are of no clinical importance can be detected because the likelihood of getting a result by chance decreases with more subjects. This is typical of large database studies with thousands of patients. Here the integrity and knowledge of the researcher and the scepticism of the reader is tested in determining if the statistically significant finding is clinically meaningful. So while we care about statistical significance from the point of view of how certain we are that the results did not occur by chance alone, we must keep our eye on the target – which is whether the difference is clinically meaningful.

A researcher who performs a study without a sample size estimate but finds a p value < .05 may say: "Well, I got a significant p value, so that must mean I had a large enough sample size "The problem with this is, just as we discussed above, if a study is under-powered, it will only be able to detect a large effect. Should an under-powered study happen to find a significant difference, it is likely that effect is inflated (i.e. the difference found in the study is higher than the actual difference) and the results are less likely to be reproducible. [2,3]

Although we often think of statistical significance as a black and white issue, and tend to completely ignore any results that are not significant, there is much more to interpreting results than that. Moreover, many statisticians will argue that the sample size calculation should not be put in the Methods section of the final paper, as there are other ways for the reader to determine if the author has reached the required sample size. This leads us to a later instalment on confidence intervals.

References

1. Wasserstein RL, Lazar NA. The ASA's statement on p-values: context, process, and purpose. *Am Stat* 2016;70:129-33.

2 Button KS, Ioannidis JP, Mokrysz C, et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci* 2013;14:365-76.

3. Ioannidis JP. Why most discovered true associations are inflated. *Epidemiology* 2008;19:640-8

Acknowledgments:
The authors would like to acknowledge Dr. Jason Oke for his helpful input.